Text-to-Video & Evaluation of Generative Models

Lecture 12

18-789

Two-Stage Training

- Stage I: Train a “Variational” Auto-Encoder, typically with GAN
  - GAN loss + Reconstruction loss + KL loss
  - It’s a “V”AE because the KL loss has a very small β (typically < 0.01)
  - 1% “Variational”, 99% Auto-encoder

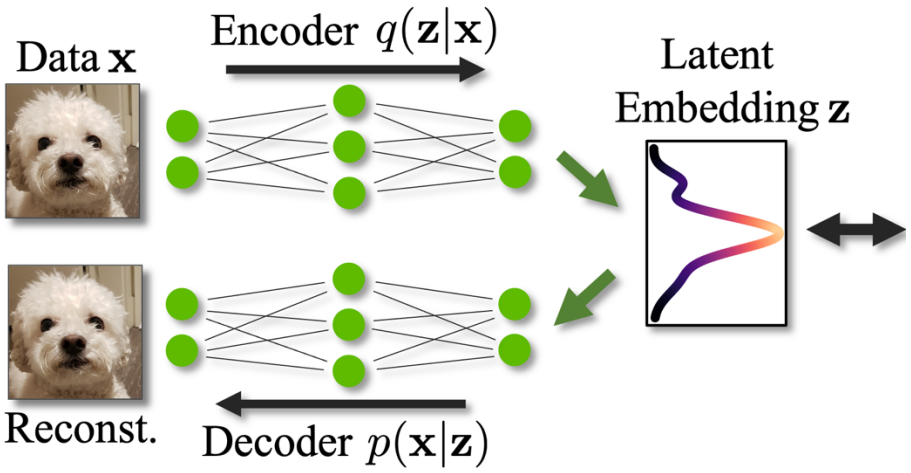


Figure source: “Score-based Generative Modeling in Latent Space”

Two-Stage Training

- Stage I: Train a "Variational" Auto-Encoder, typically with GAN
  - GAN loss + Reconstruction loss + KL loss
  - It's a "V"AE because the KL loss has a very small β (typically < 0.01)
  - 1% "Variational", 99% Auto-encoder
- Stage II: Train a Diffusion Model in the latent space

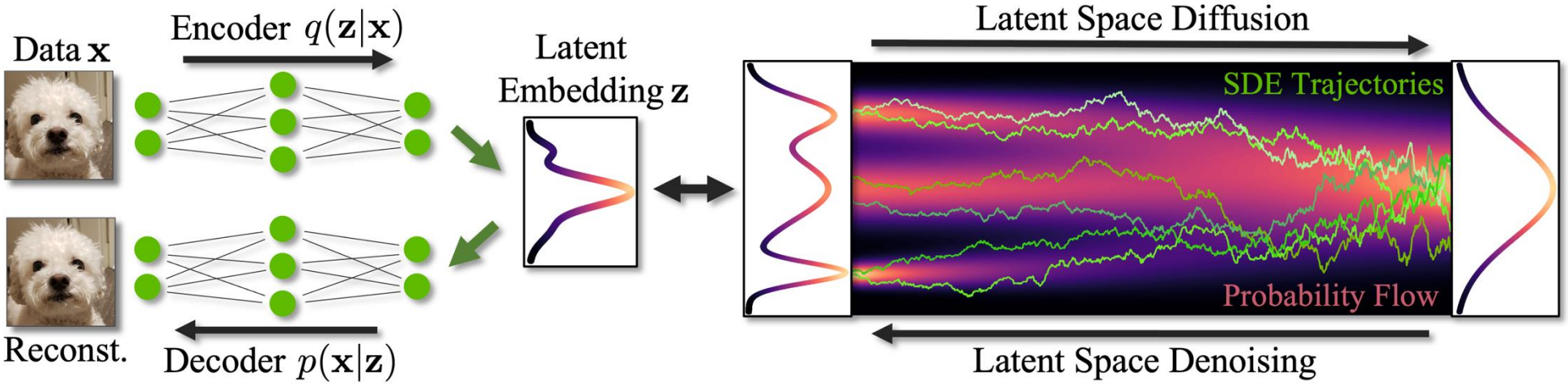
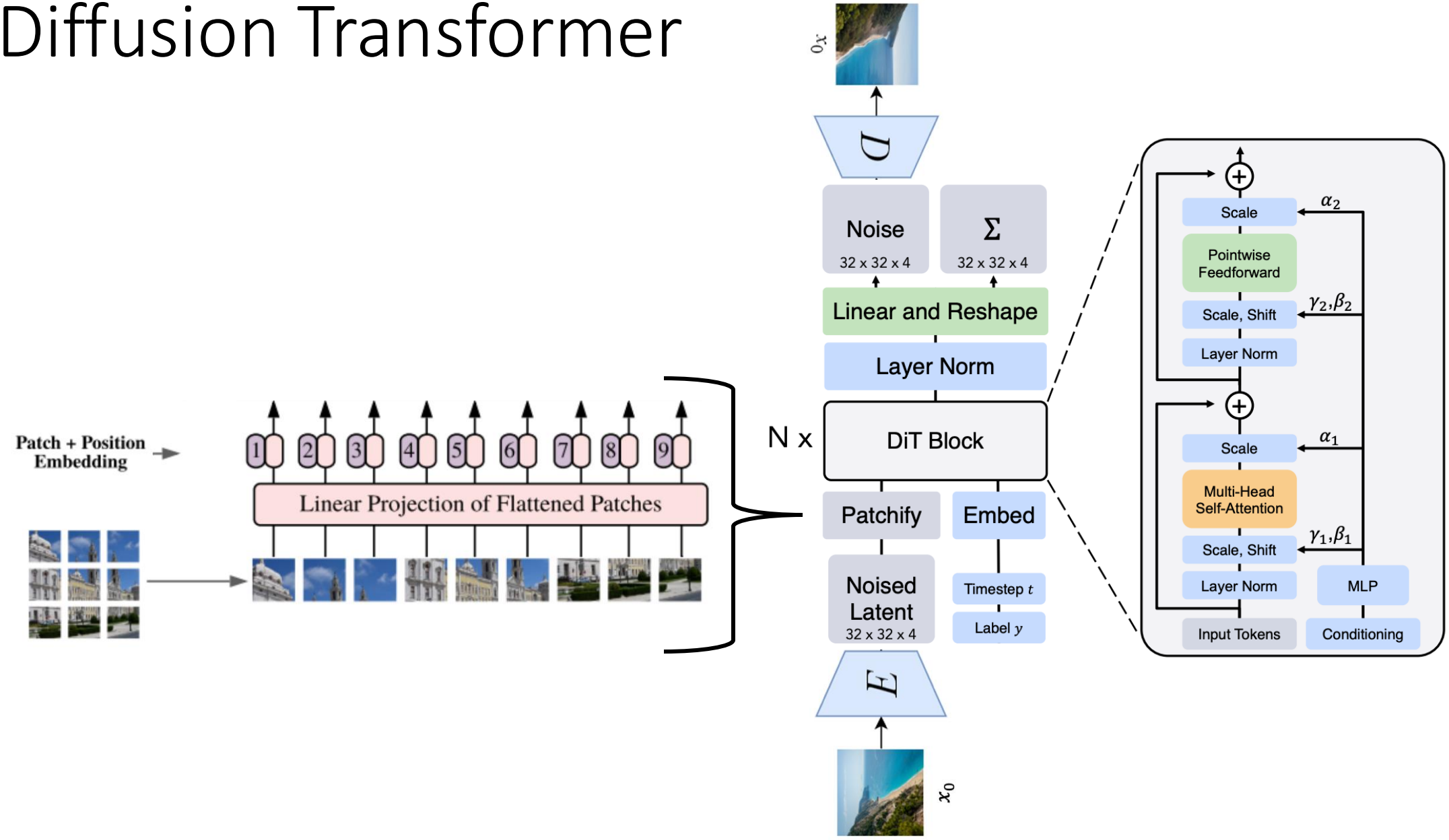
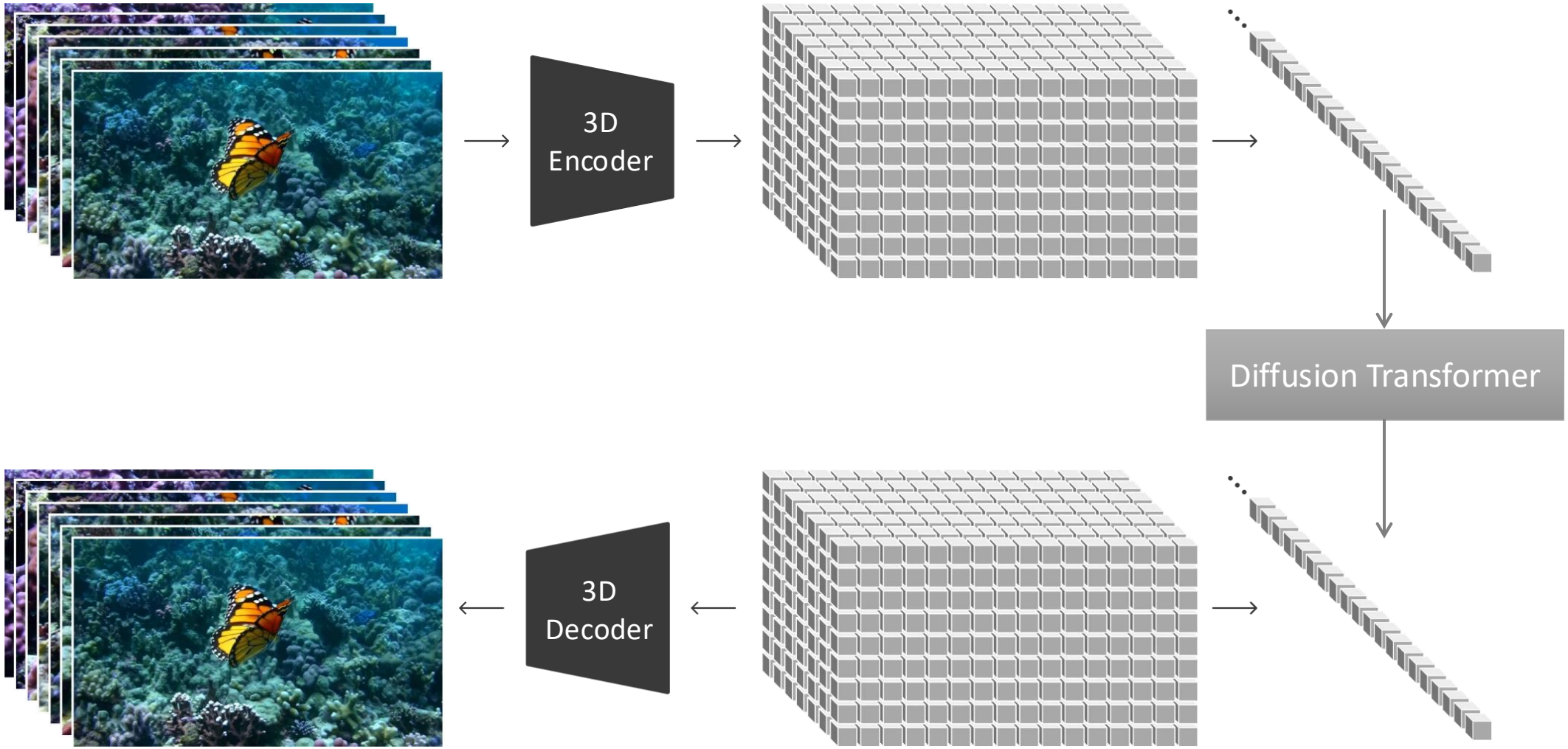


Figure source: "Score-based Generative Modeling in Latent Space"

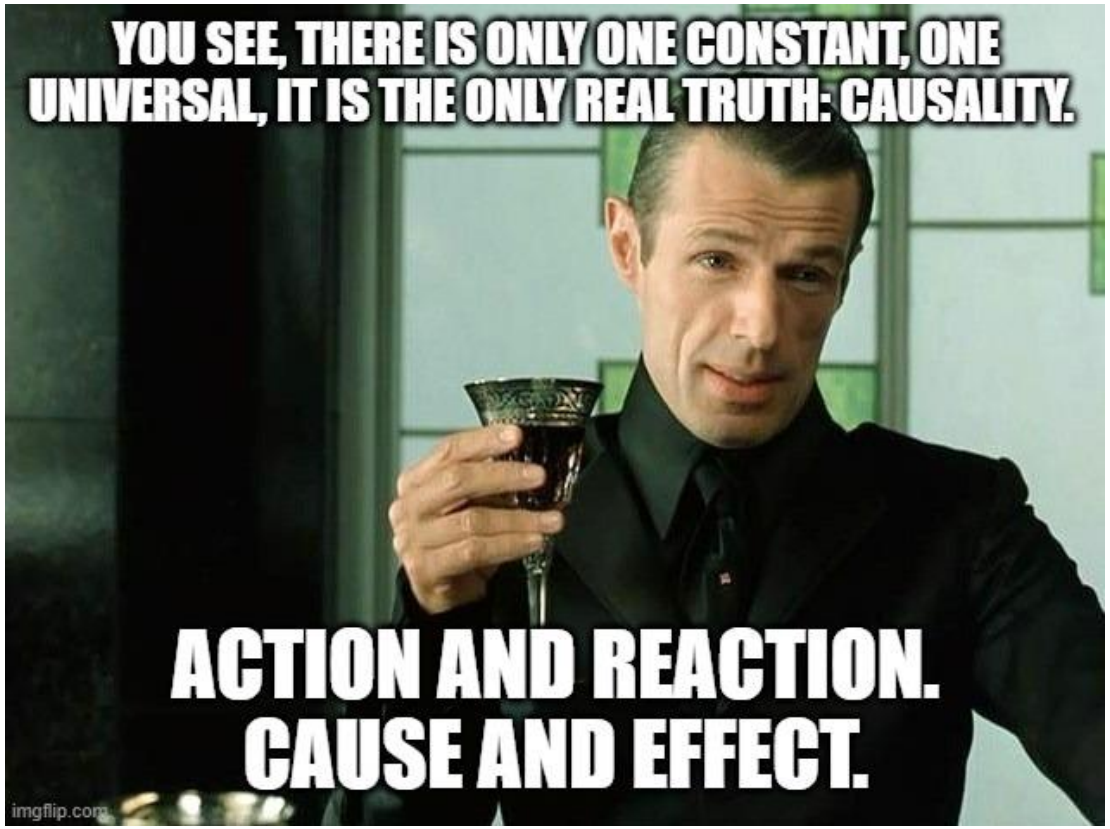
Diffusion Transformer



Video Diffusion Models



“Causality” in Video Generation



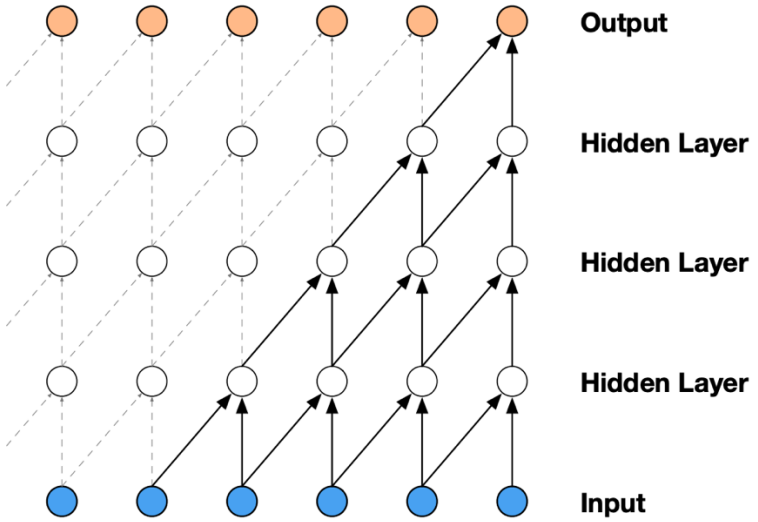
Past/Cause

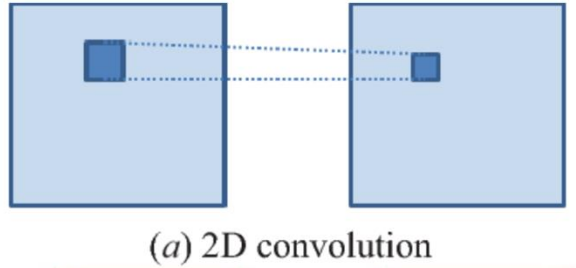
Future/Effect

Time

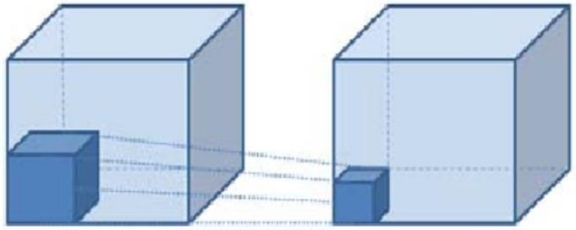
Temporally Causal “V”AE

- Each latent frame only depends on past video frames, but not future ones.
- Benefits:
  - Easier to encode/decode videos of various lengths.
  - Reuse image VAE/data.
- Causal 3D Convolution





(*a*) 2D convolution

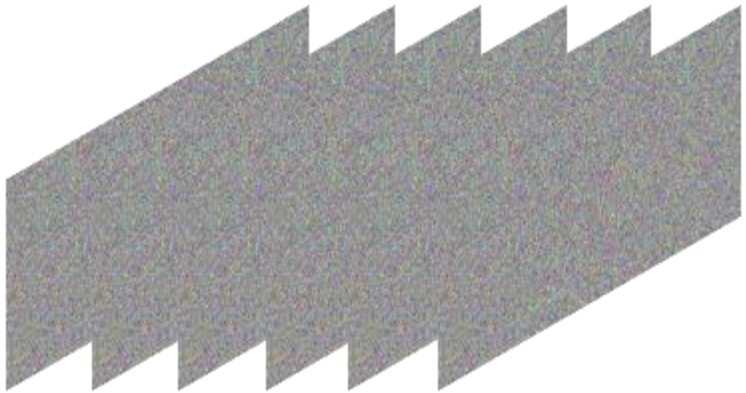


(*b*) 3D convolution

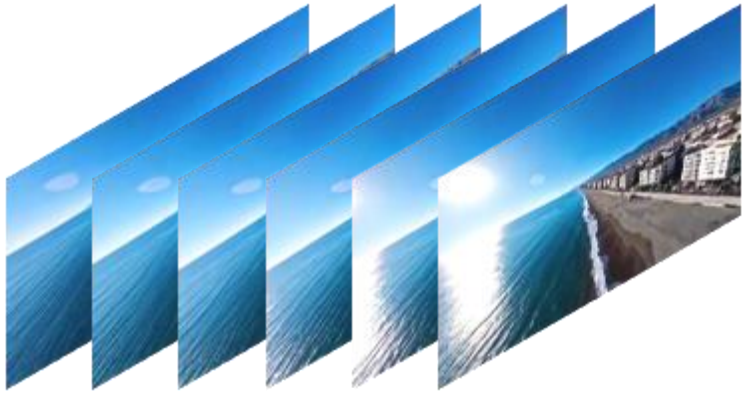
Causal Video Diffusion Models

- People have been using causal VAE to produce latent but use non-causal diffusion models to generate latent… (e.g., Sora).

Text +

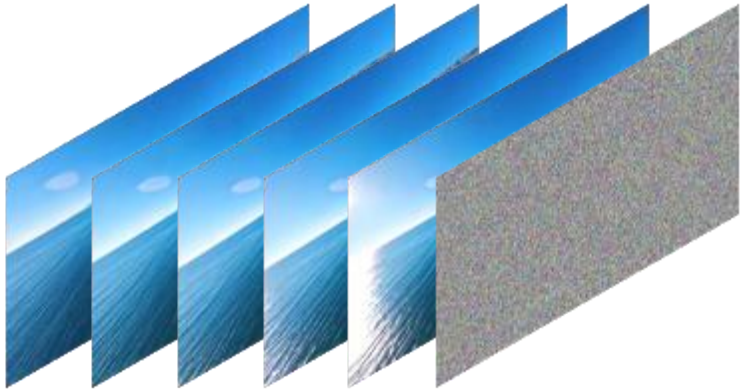


Diffusion Transformer
with **Bidirectional** Attention



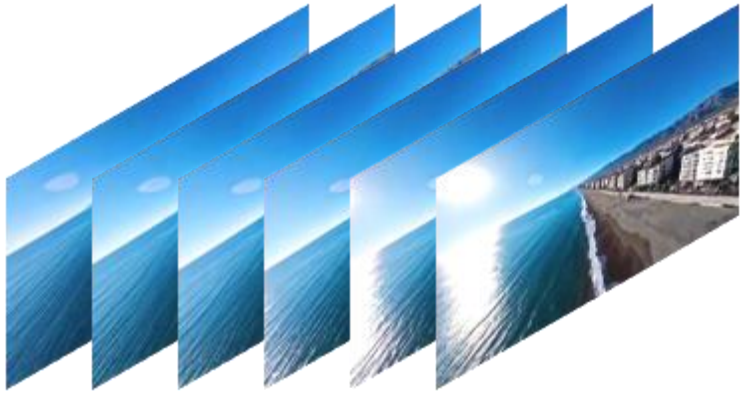
The initial latency scales **quadratically**
with the video length

Text +



.....

Autoregressive Diffusion Transformer
with **Causal** Attention



.....

The initial wait time is a **negligible constant,**
does not depend on video length.

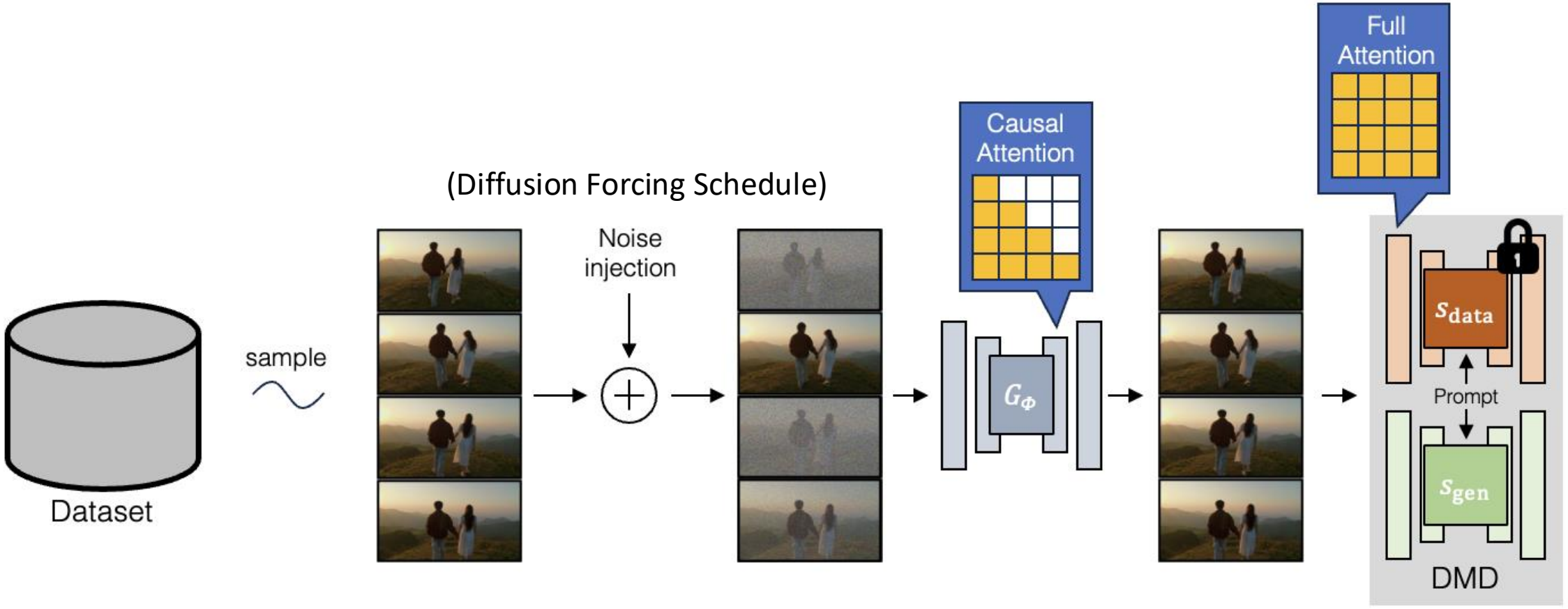
CausVid: Autoregressive Diffusion Video Generation

- We talked about Diffusion Forcing, but it
  - only have toy video experiments, based on RNN
  - still needs tens of steps to generate one frame, far from real-time

- Our work shows
  - state-of-the-art text-to-video based on Causal Transformer
  - real-time, interactive video generation, on a single GPU

Yin et al., “From Slow Bidirectional to Fast Autoregressive Video Diffusion Models”, CVPR 2025

CausVid Algorithm



DMD = Distribution Matching Distillation

Bidirectional Teacher

Preparing...

Progress: 0/1

```
16] -1/-1/-1->0->-1 [17] -1/-1/-1->0->-1 [18] -1/-1/-1->0->-1 [19] -1/-1/-1->0->-1 [20] -1/-1/-1
->0->-1 [21] -1/-1/-1->0->-1 [22] -1/-1/-1->0->-1 [23] -1/-1/-1->0->-1 [24] -1/-1/-1->0->-1 [25]
 -1/-1/-1->0->-1 [26] -1/-1/-1->0->-1 [27] -1/-1/-1->0->-1 [28] -1/-1/-1->0->-1 [29] -1/-1/-1->0
->-1 [30] -1/-1/-1->0->-1 [31] -1/-1/-1->0->-1
xuhuang-0464291930-0-0:504194:506058 [0] NCCL INFO P2P Chunksize set to 131072
xuhuang-0464291930-0-0:504194:506058 [0] NCCL INFO Connected all rings
xuhuang-0464291930-0-0:504194:506058 [0] NCCL INFO Connected all trees
xuhuang-0464291930-0-0:504194:506058 [0] NCCL INFO 32 coll channels, 32 collnet channels, 0 nvls
 channels, 32 p2p channels, 32 p2p channels per peer
xuhuang-0464291930-0-0:504194:506058 [0] NCCL INFO TUNER/Plugin: Failed to find ncclTunerPlugin_
v2, using internal tuner instead.
xuhuang-0464291930-0-0:504194:506058 [0] NCCL INFO ncclCommInitRank comm 0x55a809dbc940 rank 0 n
ranks 1 cudaDev 0 nvmlDev 0 busId 53000 commId 0x54a120ee0c121fbc - Init COMPLETE
```

CausVid (Ours)

Preparing...

Progress: 0/15

00:00.0

```
xuhuang-0464291930-0-0:472752:473617 [0] NCCL INFO NET/OFI NIC group 5 device #3 0000:a4:00.0
xuhuang-0464291930-0-0:472752:473617 [0] NCCL INFO NET/OFI NIC group 6 device #0 0000:b8:00.0
xuhuang-0464291930-0-0:472752:473617 [0] NCCL INFO NET/OFI NIC group 6 device #1 0000:b7:00.0
xuhuang-0464291930-0-0:472752:473617 [0] NCCL INFO NET/OFI NIC group 6 device #2 0000:b6:00.0
xuhuang-0464291930-0-0:472752:473617 [0] NCCL INFO NET/OFI NIC group 6 device #3 0000:b5:00.0
xuhuang-0464291930-0-0:472752:473617 [0] NCCL INFO NET/OFI NIC group 7 device #0 0000:c9:00.0
xuhuang-0464291930-0-0:472752:473617 [0] NCCL INFO NET/OFI NIC group 7 device #1 0000:c8:00.0
xuhuang-0464291930-0-0:472752:473617 [0] NCCL INFO NET/OFI NIC group 7 device #2 0000:c7:00.0
xuhuang-0464291930-0-0:472752:473617 [0] NCCL INFO NET/OFI NIC group 7 device #3 0000:c6:00.0
xuhuang-0464291930-0-0:472752:473617 [0] NCCL INFO NET/OFI Libfabric provider associates MRs w
ith domains
xuhuang-0464291930-0-0:472752:473617 [0] NCCL INFO Using non-device net plugin version 0
xuhuang-0464291930-0-0:472752:473617 [0] NCCL INFO Using network AWS Libfabric
```

Vbench Evaluation

Model Name (clickable)	Sampled by	Evaluated by	Accessibility	Date	Total Score
Wan2.1(2025-02-24)	Wan Team	VBench Team	API	2025-02-24	86.22%
IPOC	IPOC Team	VBench Team	Close Source	2025-02-28	85.71%
MiracleVision V5	MVV Team	VBench Team	API	2025-01-21	85.23%
Wan2.1	Wan Team	VBench Team	API	2025-01-08	84.70%
Sora	VBench Team	VBench Team	API	2025-01-14	84.28%
CausVid(2025-01-02 5s)	CausVid Team	VBench Team	Close Source	2025-01-02	84.27%
CausVid	CausVid Team	VBench Team	Close Source	2024-12-07	83.88%
Luma	VBench Team	VBench Team	API	2025-01-14	83.61%
EasyAnimateV5.1	EasyAnimate Team	VBench Team	Open Source	2025-01-22	83.42%
MiniMax-Video-01	VBench Team	VBench Team	API	2024-10-01	83.41%
STIV (Apple)	Apple Team	VBench Team	Close Source	2024-12-19	83.35%
HunyuanVideo (Open-Source	VBench Team	VBench Team	Open Source	2024-12-16	83.24%
Gen-3 (2024-07)	VBench Team	VBench Team	API	2024-07-25	82.32%
Vchitect-2.0 (VEnhancer)	VBench Team	VBench Team	Open Source	2024-09-20	82.24%
CogVideoX1.5-5B (5s SAT pr	VBench Team	VBench Team	Open Source	2024-11-15	82.17%

10-200x faster than other methods on the table!

Interactive “World Model”

Evaluation of Generative Models

Training

cat

Generative Model

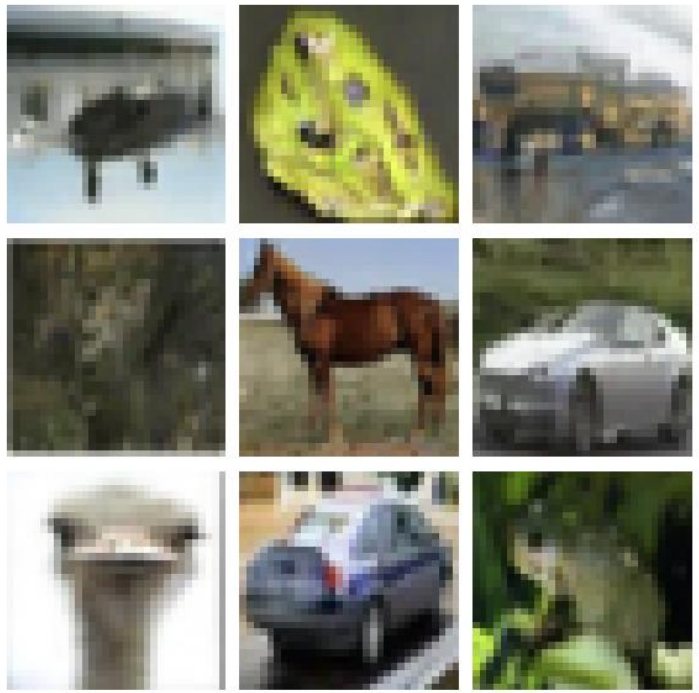
How to evaluate the performance of our model?

Log-likelihood

- $\frac{1}{N}\sum_{i=1}^{N}\log p_\theta(x_i)$ computed on **test set** $\{x_i\}$
- Similar to "perplexity" in language modeling
  - For a sentence $\boldsymbol{x} = (x^1, x^2, \ldots, x^t)$
  - $\log \mathrm{PPL}(x^1, x^2, \ldots, x^t) = -\frac{1}{t}\sum_{s=1}^{t}\log p_\theta(x^s|x^{<s})$
- Problems:
  - Not all models have tractable likelihoods (e.g., VAEs, GANs)
  - Does not correlate well with quality

Log-likelihood does not correlate well with quality

PixelCNN
(Autoregressive)



DDPM
(Diffusion)

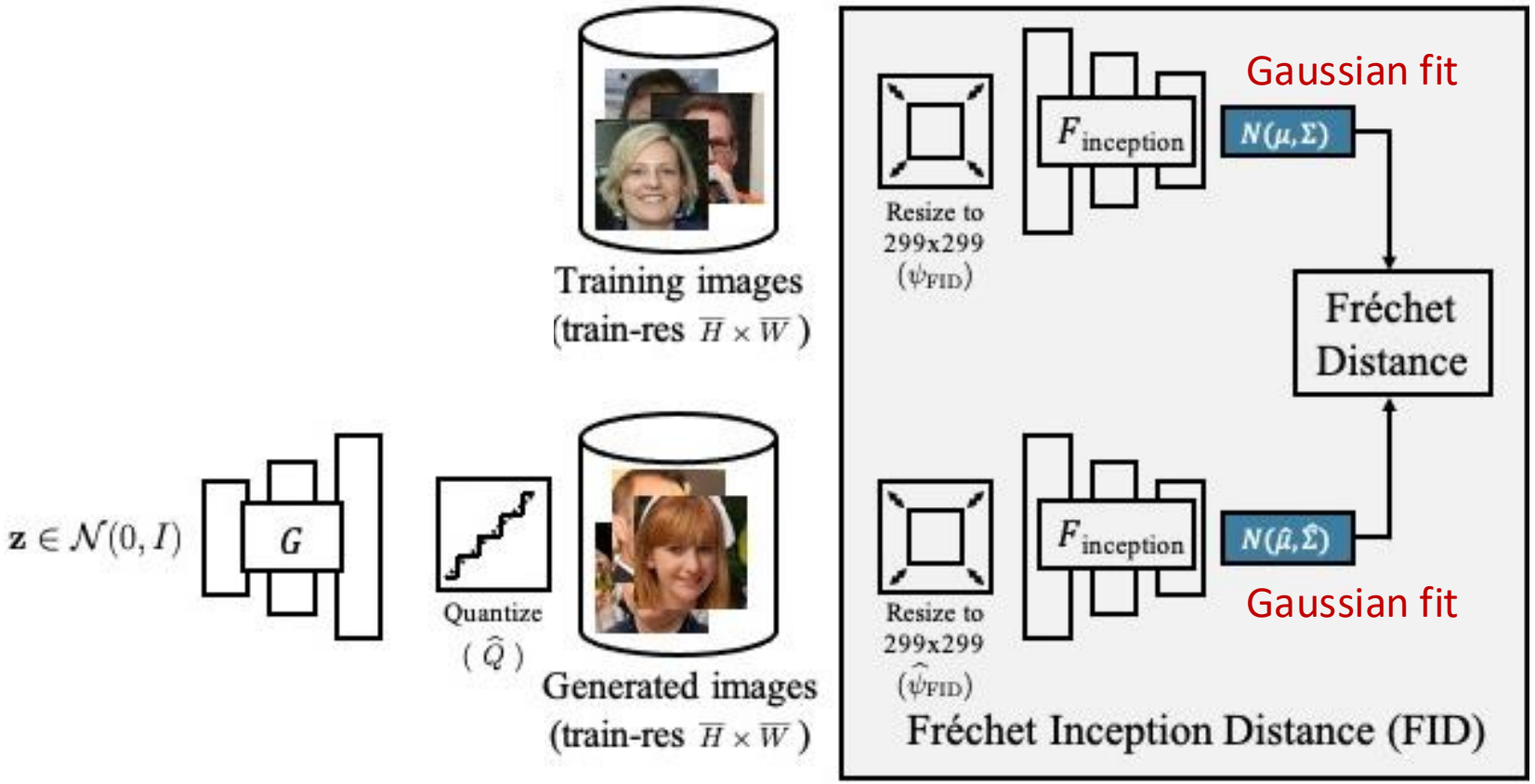
Table 2: NLLs and FIDs (ODE) on CIFAR-10.

Model	NLL Test ↓	FID ↓
RealNVP (Dinh et al., 2016)	3.49	-
iResNet (Behrmann et al., 2019)	3.45	-
Glow (Kingma & Dhariwal, 2018)	3.35	-
MintNet (Song et al., 2019b)	3.32	-
Residual Flow (Chen et al., 2019)	3.28	46.37
FFJORD (Grathwohl et al., 2018)	3.40	-
Flow++ (Ho et al., 2019)	3.29	-
DDPM (L) (Ho et al., 2020)	$\leqslant 3.70^*$	13.51
DDPM (L_{simple}) (Ho et al., 2020)	$\leqslant 3.75^*$	3.17
PixelCNN (Oord et al., 2016)	3.03	
PixelCNN++ (Salimans et al., 2017)	2.92	
Image Transformer (Parmar et al., 2018	2.90	
PixelSNAIL (Chen et al., 2017)	2.85	
Sparse Transformer 59M (strided)	**2.80**	
DDPM	3.28	3.37
DDPM cont. (VP)	3.21	3.69
DDPM cont. (sub-VP)	3.05	3.56
DDPM++ cont. (VP)	3.16	3.93
DDPM++ cont. (sub-VP)	3.02	3.16
DDPM++ cont. (deep, VP)	3.13	3.08
DDPM++ cont. (deep, sub-VP)	**2.99**	**2.92**

Measure distribution similarity in *perceptual space*

- Log-likelihood = $-\mathrm{KL}(p_{data}||p_{\theta})$
- We still want to measure distribution similarity, but in a space that's more perceptual (instead of in pixel space)

Fréchet inception distance (FID)

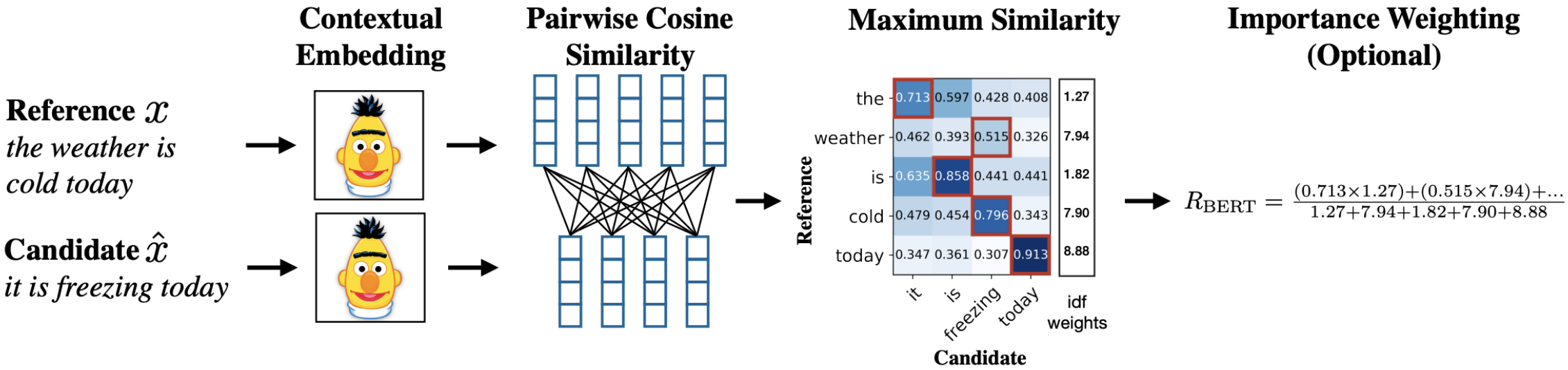


Wasserstein distance between Gaussians

$$\text{FID} = \left\|\mu_{\text{real}} - \mu_{\text{gen}}\right\|^2 + \text{Tr}(\Sigma_{\text{real}} + \Sigma_{\text{gen}} - 2(\Sigma_{\text{real}}\Sigma_{\text{gen}})^{0.5})$$

Kynkäänniemi et al., “GANs trained by a two time-scale update rule converge to a local nash equilibrium”, NeurIPS 2017

BERTScore

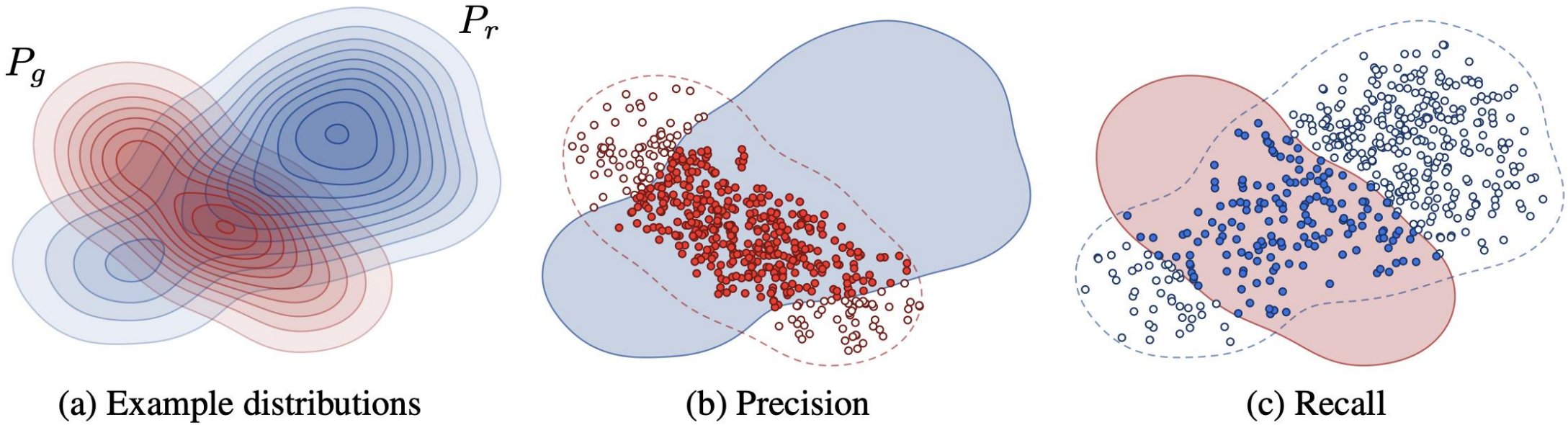


Zhang et al., "BERTScore: Evaluating Text Generation with BERT", ICLR 2020

Multi-axis Evaluation

- Quality/Precision
- Diversity/Recall
- Alignment (in conditional generation)

Precision/Recall

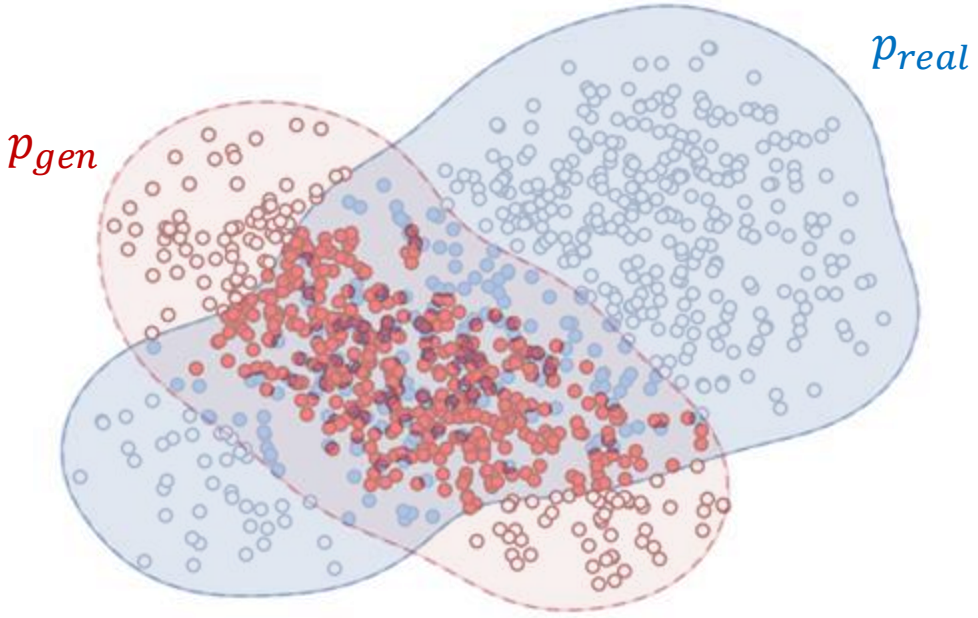


(a) Example distributions (b) Precision (c) Recall

Kynkäänniemi et al., “Improved Precision and Recall Metric for Assessing Generative Models”, NeurIPS 2019

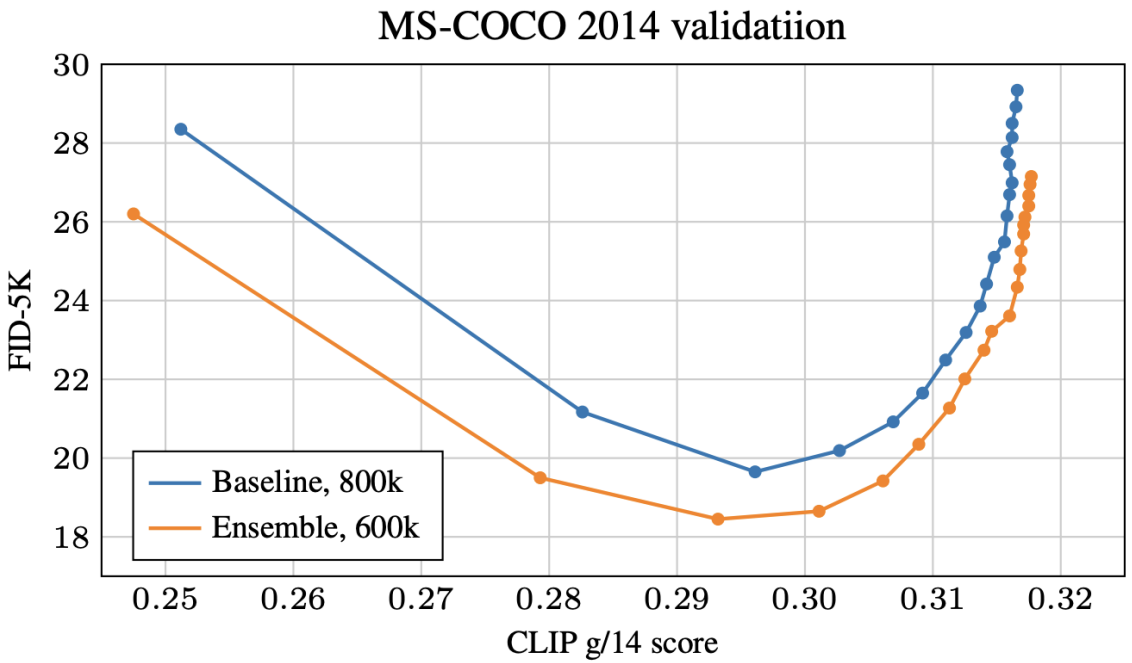
Precision/Recall

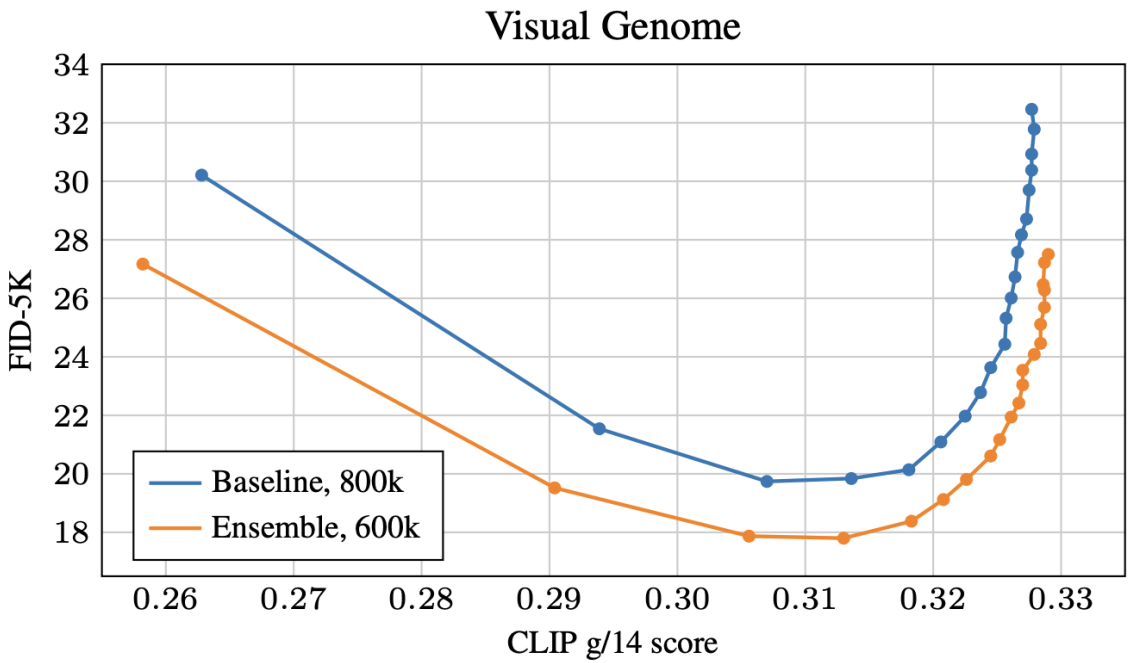
- Mix real and generated data, and compute k-NN for each sample
- If a real data has a generated data in its k-NN, it's "recalled"
- If a generated data has a real data in its k-NN, it's "precise"



Text-to-Image Evaluation

- Quality/Diversity: Measure by FID
- Text Alignment: Measure by CLIP score: $\cos(\text{CLIP}(\text{text}), \text{CLIP}(\text{image}))$





Human Preference

TEXT TO IMAGE ARENA

0/30 to view your model preferences 👀

+ Submit prompt

Try the new Speech Arena

Which image best reflects this prompt?

Steampunk workshop with time travelers and eccentric machinery, portrayed in cartoon style.

Prefer (← Key)

Prefer (→ Key)

Elo Rating System

If players A, B have ratings R_A and R_B, the expected score of players is

$$E_A = \frac{1}{1+10^{(R_B-R_A)/400}} \qquad E_B = \frac{1}{1+10^{(R_A-R_B)/400}}$$

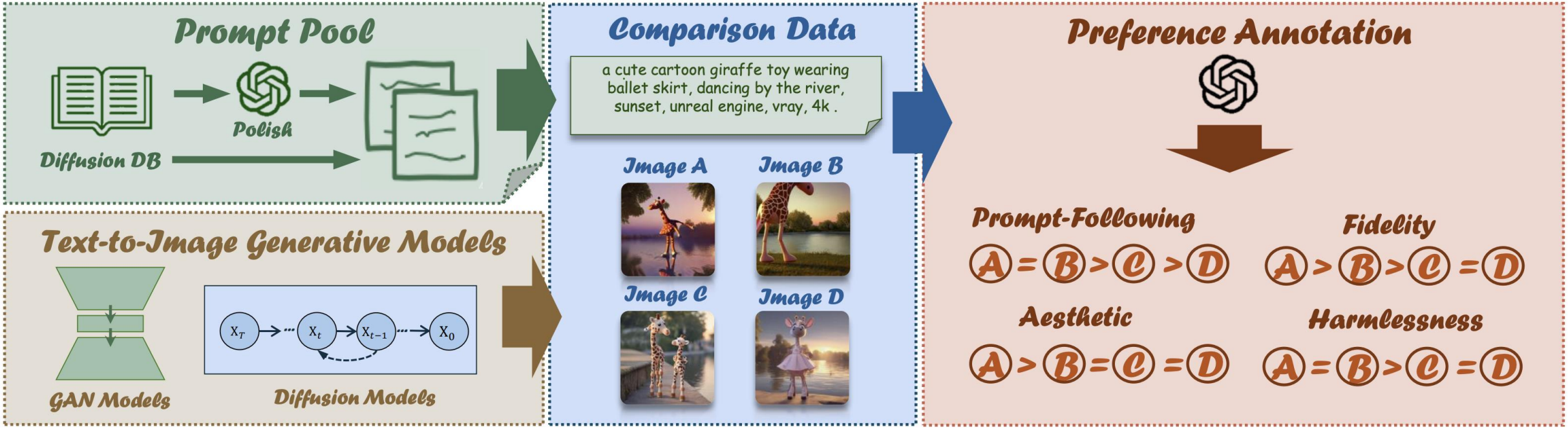
After the game, players actually score S_A, S_B so their rating is updated

$$R'_A = R_A + K(S_A - E_A) \qquad R'_B = R_B + K(S_B - E_B)$$

where K is the maximum possible rating gain or loss per match

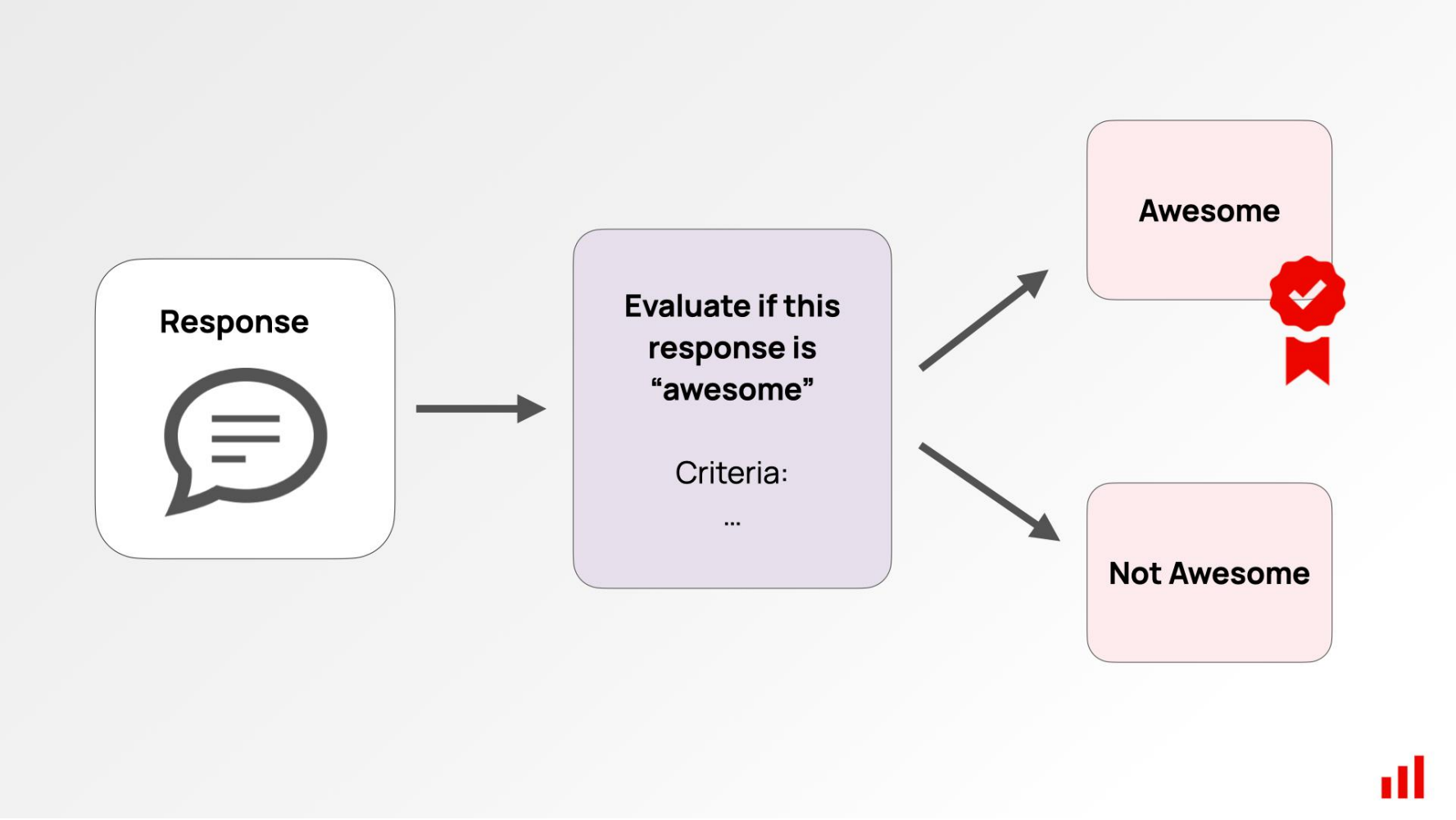
Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization	License
1	2	Grok-3-Preview-02-24	1406	+8/-6	9109	xAI	Proprietary
1	1	GPT-4.5-Preview	1400	+5/-6	8596	OpenAI	Proprietary
3	6	Gemini-2.0-Flash-Thinking-Exp-01-21	1383	+6/-4	21124	Google	Proprietary
3	3	Gemini-2.0-Pro-Exp-02-05	1380	+4/-4	19038	Google	Proprietary
3	2	ChatGPT-4o-latest (2025-01-29)	1375	+6/-4	20936	OpenAI	Proprietary
6	4	DeepSeek-R1	1360	+7/-5	11507	DeepSeek	MIT
6	10	Gemini-2.0-Flash-001	1355	+4/-5	16845	Google	Proprietary
6	3	o1-2024-12-17	1352	+4/-6	23441	OpenAI	Proprietary
8	10	Gemma-3-27B-it	1340	+8/-8	5028	Google	Gemma
9	10	Qwen2.5-Max	1339	+4/-5	15607	Alibaba	Proprietary
9	7	o1-preview	1335	+4/-4	33187	OpenAI	Proprietary
11	10	o3-mini-high	1326	+6/-5	12773	OpenAI	Proprietary
12	13	DeepSeek-V3	1318	+4/-3	22857	DeepSeek	DeepSeek
12	17	QwQ-32B	1312	+10/-10	2732	Alibaba	Apache 2.0
13	11	Command A (03-2025)	1313	+8/-8	2771	Cohere	CC-BY-NC-4.0
13	15	Qwen-Plus-0125	1310	+6/-8	6058	Alibaba	Proprietary

Replace Human with Vision-Language Models



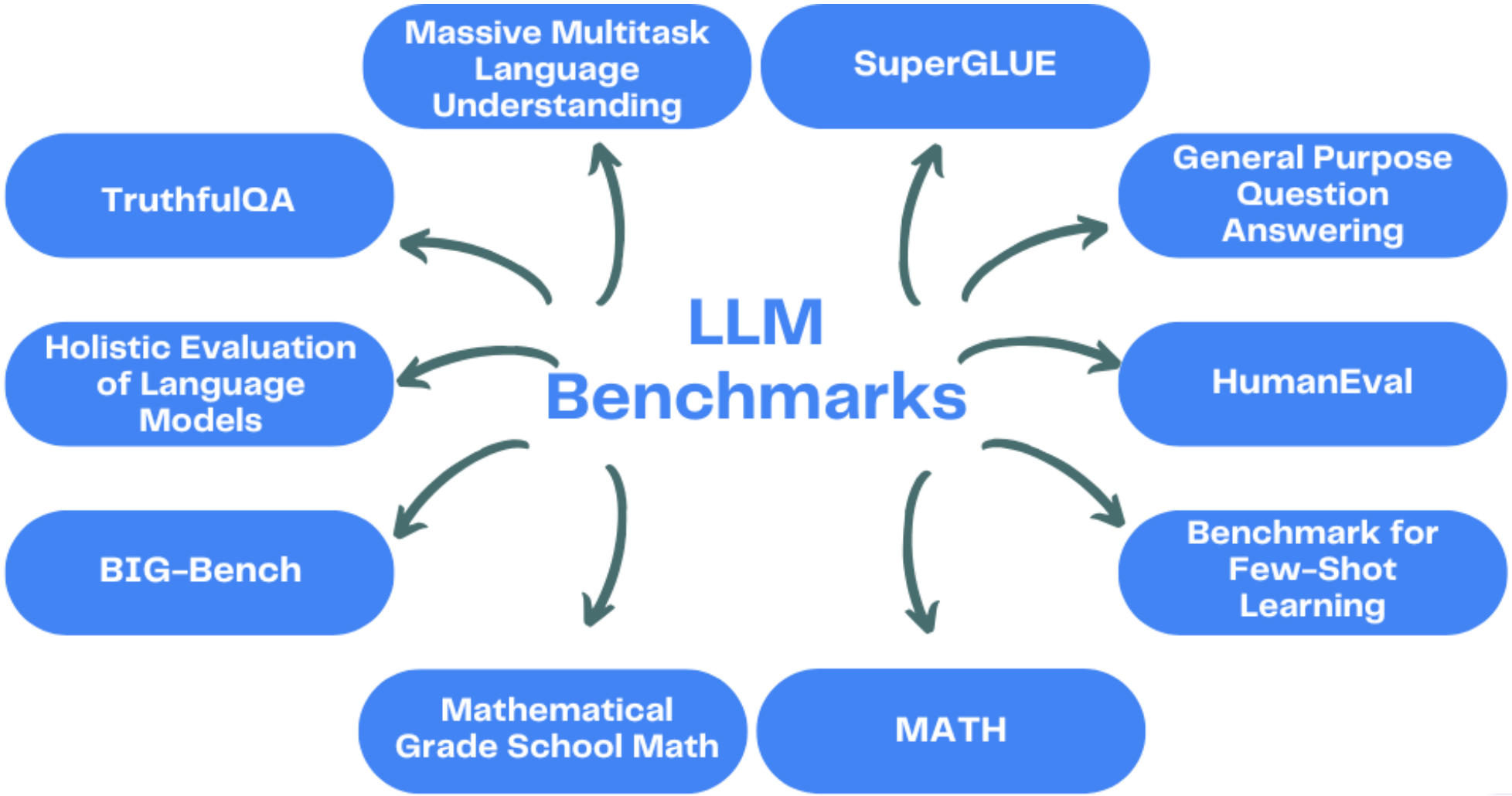
Wu et al., “Multimodal Large Language Model is a Human-Aligned Annotator for Text-to-Image Generation”, ECCV 2024

Evaluating LLMs: LLM-as-a-judge



Evidently AI, “LLM-as-a-judge: a complete guide to using LLMs for evaluations”

Task-specific evaluation



Summary

- Evaluating Generative Models is highly non-trivial
- Log-likelihood is insufficient
- Automatic metrics can measure quality/diversity/alignment
- For accurate evaluation of general quality, human evaluation is still better. VLM/LLMs can also serve the role of human evaluators
- For LLMs, task-specific evaluation can be conducted.

Ultimately, evaluation should align with intended downstream applications.